DOCUMENT RESUME

TM 000 654

ED 052 248

AUTHOR	Temp, George
TITLE	Test Bias: Validity of the SAT for Blacks and Whites in Thirteen Integrated Institutions.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	RB-71-2; RDR-70-71-6
PUB DATE	Jan 71
NOTE	18p.
EDRS PRICE	EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS	Caucasian Students, College Admission, College Freshmen, *College Integration, *Grade Point Average, Grade Prediction, Measurement Instruments, Negro Students, Predictive Ability (Testing), *Predictive Validity, *Racial Differences, *Test Bias
IDENTIFIERS	SAT, *Scholastic Aptitude Test

ABSTRACT

Differential prediction of grade point average for black and white freshman students was empirically investigated at 13 integrated institutions by comparison of regression planes. Particular attention was given to the possibility that prediction procedures that are appropriate for white (majority) students would under-predict the performance of black (minority) students. The data tend to support, among others, the following generalizations: (1) a single regression plane cannot be used to predict freshman GPA for both blacks and whites in many of the institutions studied; (2) nevertheless, if prediction of GPA from SAT scores is based upon prediction equations suitable for majority students, then black students, as a group, are predicted to do about as well as (or better than) they actually do. Analysis demonstrated that a general conclusion applicable to all institutions is not justified. Admissions officers are urged to consider and conduct institutional self-studies routinely on the question of differential predictive validity. (Author/LR)

ED052248

COLLEGE ENTRANCE EXAMINATION BOARD
RESEARCH AND DEVELOPMENT REPORTS

RDR-70-71, NO. 6

•

RESEARCH BULLETIN
RB-71-2 JANUARY 1971

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

Test Bias: Validity of the SAT for Blacks and Whites in Thirteen Integrated Institutions

•

George Temp

TM 000 654

EDUCATIONAL TESTING SERVICE
PRINCETON, NEW JERSEY
BERKELEY, CALIFORNIA

1

ED052248

TEST BIAS: VALIDITY OF THE SAT FOR BLACKS AND WHITES IN THIRTEEN INTEGRATED INSTITUTIONS

Abstract

Differential prediction for black and white students was empirically investigated at 13 institutions by comparison of regression planes. Particular attention was given to the possibility that prediction procedures that are appropriate for white (majority) students would underpredict the performance of black (minority) students. The data tend to support, among others, the following generalizations: (1) a single regression plane cannot be used to predict freshman GPA for both blacks and whites in many of the institutions studied; (2) nevertheless, if prediction of GPA from SAT scores is based upon prediction equations suitable for majority students, then black students, as a group, are predicted to do about as well as (or better than) they actually do.

Test Bias: Validity of the SAT for Blacks and Whites

in Thirteen Integrated Institutions

George Temp[1]

Educational Testing Service

Earlier work by Cleary (1968), Stanley and Porter (1967), and others (see references in those two articles) has led informed observers (Kendrick & Thomas, 1970) to conclude that the College Board Scholastic Aptitude Test (SAT) and other such tests do predict at the usual levels for minority black students. However, persistent charges that admissions testing, in general, and the SAT, in particular, are a part of the college admissions problem for minority black students indicated the need to examine further empirical evidence.

It is apparent that college programs for blacks are expanding rapidly and greater numbers of students with lower SAT scores are being admitted. Such changes might affect the predictive validity of the SAT in certain institutions. The present study was not an investigation of admissions practices at various institutions, but an examination of the predictive validity of the SAT for blacks and whites at 13 integrated institutions.

Selection of Schools

A list of integrated colleges and universities was taken from a federal survey of Negro undergraduate enrollment (Chronicle, 1969). Institutions were selected and invited to participate in the study if

---

[1]The author is indebted to Dr. Abraham Carp for critical support and guidance throughout the conduct of this study and the preparation of the report. Dr. William H. Angoff and Dr. Robert L. Linn offered constructive and helpful criticisms of an earlier draft of this paper for which the author is grateful. Dr. John Bianchini provided the computer program for the Gulliksen/Wilks regression analyses.

(1) they reported that they enrolled 300 or more blacks;

(2) they indicated in College Board publications that they required SAT scores of applicants for admission.

In addition, eight "selective" and/or "innovative" institutions with less than 300 blacks were invited to participate. The final list of 27 institutions was, of course, not representative of all U.S. colleges, but represented a number of case studies of predictive validity with particular institutions. Although almost unanimous institutional cooperation was received in the initial phases of the study, several colleges withdrew for various reasons during the course of the year.[2] The final number of participating colleges was 13. Eight of these had 300 or more blacks in undergraduate status and five of the "innovative" institutions had a smaller number of black students.

Procedures of Participation

The cooperating institutions were charged with the responsibility of supplying the data, which consisted of

(a) SAT-verbal and SAT-math (SAT-V and M) scores;

(b) grade point average (GPA) computed at the end of the freshman year or at the point of withdrawal from the institution.

The groups for which the above data were requested were defined as follows:

Group 1: Black students registered as fulltime freshmen during the 1968-69 academic year, preferably in a single curriculum area (e.g., liberal arts).

---

[2]The most common reason was an inability to identify black students from the data readily available.

Group 2: A sample of registered white students from the same class and curriculum area defined above, drawn randomly until their number equaled that of Group 1.

Institutions returned a roster for Group 1 (blacks) and for Group 2 (whites) following the format shown below.

Typical Data Format[a]

SAT-V	SAT-M	End of Year Status	Attempted Credits	GPA
378	415		45	2.35
275	300	Probation	40	1.00

[a]Fictitious data

Analysis Procedures

All data were analyzed by a procedure that utilizes the rationale presented by Gulliksen and Wilks (1950). This procedure tests the hypothesis that the regression systems in two (or more) groups may be regarded as being essentially the same. Specifically, the homogeneity of the errors of estimate, the slopes, and the intercepts of the regression equations are tested sequentially to determine whether or not the same plane can be applied to both blacks and whites at each college in the study.

Potthoff (1966) has presented a detailed review of the statistical aspects of the technical problem of test bias. His definition states that "a test is not biased if individuals from different groups who have the same test scores also have the same expected criterion score." This suggests the methodology for examination of test bias is detailed comparison of regression planes and

prediction equations. In particular, underprediction of black students' actual performance would be evidence of test bias.

Actually, the interpretation of the evidence is not quite that direct. It is possible, for instance, that the predictor measurement is less reliable for one group than for another, thus accounting for some of the over- or underprediction (Linn & Werts, 1971). It is also possible that one group may be graded more rigorously or more leniently depending upon circumstances and situations independent of the predictor instrument. Rock (1970) has described many possible sources of bias in any prediction system and indicates the complexity of interpretations of evidence in this area.

It was anticipated that the Gulliksen-Wilks tests would generally indicate that a single regression plane was not appropriate in this study. Accordingly, it was planned to analyze each set of data in another way. The second analysis involved application of the regression system derived from the majority (or white) group data, to make a prediction for the average score of the black group. Predicted performance was then compared with actual performance of the black group studied at each institution.

Regression Tests for the Two Groups

The first step in testing whether or not the same regression plane could apply to blacks and whites was to determine if the standard errors of estimate of grade point average from the predictor were essentially the same. (The SAT was used as a predictor in four different ways. First, the verbal portion alone; second, the math portion alone; third, the verbal plus math as a single composite score; and fourth, verbal plus math in the best weighted multiple regression. The rationale for this was that institutions might vary on how

much of the SAT they used and whether or not they bothered to calculate a weighted multiple. (At this point it might be well to note that only the influence of the test was under investigation. That is, no attempt was made to find the best total multiple prediction system for each college, or to examine actual admissions policies and procedures.)

The second test in the Gulliksen-Wilks procedure is a test of equality of slopes, assuming that the standard errors of estimate are the same.

The final test is whether or not the intercepts of the y axis are essentially the same, given that the slopes and errors of estimate are equal.

To summarize, the testing procedure is sequential, and significant results at any stage indicate that the hypothesis of a common regression plane is not tenable. Table 1 presents the results of those tests.

It is readily seen from Table 1 that the regression planes for blacks and whites cannot be considered to be the same. All except nine of the 52 regressions were significantly different at some point of the comparison. Institutions are cautioned to examine carefully their own experience with prediction of freshman grade point average for different subgroups. It is very likely on the basis of these data that a common prediction system is not possible and that a separate prediction system would have to be developed for each subgroup.

Prediction of Minority Performance from Majority Regression

Those who have challenged the use of admissions tests have not only maintained that different prediction systems are technically required by the available evidence, but also that a prediction system based upon majority (white) students would penalize black applicants by underpredicting their college performance. The effect of underprediction might be to deny admission to some capable minority students.

Table 1

Gulliksen-Wilks Tests of Equality of Regression Planes: Do the Regressions Significantly Differ? $P < .05$

College	GPA on...	on error of estimate	on slopes	on intercepts	Is a single regression plane tenable?
1	SAT Verbal			yes	no
	Math			yes	no
	V+M			yes	no
	V,M			yes	no
2	SAT Verbal			yes	no
	Math			yes	no
	V+M			yes	no
	V,M			yes	no
3	SAT Verbal	yes	n.a.	n.a.	no
	Math	yes	n.a.	n.a.	no
	V+M	yes	n.a.	n.a.	no
	V,M	yes	n.a.	n.a.	no
4	SAT Verbal			yes	no
	Math			yes	no
	V+M			yes	no
	V,M			yes	no
5	SAT Verbal	yes	n.a.	n.a.	no
	Math		yes	n.a.	no
	V+M	yes	n.a.	n.a.	no
	V,M	yes	n.a.	n.a.	no
6	SAT Verbal			yes	no
	Math			yes	no
	V+M				yes
	V,M				yes
7	SAT Verbal			yes	no
	Math			yes	no
	V+M			yes	no
	V,M			yes	no
8	SAT Verbal				yes
	Math			yes	no
	V+M				yes
	V,M				yes
9	SAT Verbal	yes	n.a.	n.a.	no
	Math	yes	n.a.	n.a.	no
	V+M	yes	n.a.	n.a.	no
	V,M	yes	n.a.	n.a.	no
10	SAT Verbal	yes	n.a.	n.a.	no
	Math	yes	n.a.	n.a.	no
	V+M	yes	n.a.	n.a.	no
	V,M	yes	n.a.	n.a.	no
11	SAT Verbal		yes	n.a.	no
	Math			yes	no
	V+M		yes	n.a.	no
	V,M		yes	n.a.	no
12	SAT Verbal				yes
	Math				yes
	V+M				yes
	V,M				yes
13	SAT Verbal	yes	n.a.	n.a.	no
	Math			yes	no
	V+M	yes	n.a.	n.a.	no
	V,M	yes	n.a.	n.a.	no

Note: n.a. = not applicable. The tests are sequential and one yes indicates a significantly different dimension of the regression plane.

The charge that a majority prediction system penalizes black applicants is not directed at the well-known error of estimate. Instead, the charge implies that some further systematic (and even greater) error is made when the minority group is predicted with the majority system. The second of the planned analyses was designed to provide evidence on this question. Predictions of the performance of the black group were made by using the majority regression equation. This prediction was then compared with the actual performance of the minority group to see what difference, if any, appeared and its direction. Table 2 reports this information.

It is evident from Table 2 that if prediction of GPA from SAT scores is based upon regression equations suitable for majority (white) students, then minority (black) students, as a group, are predicted to do about as well as (or better than) they actually do. In four cases (colleges 2, 5, 7, 10) the predicted GPA is more than 1/2 of a standard deviation above the actual GPA, and in only three cases (colleges 8, 11, 13) is the GPA less than 1/5 of a standard deviation above the actual GPA. In one instance there is an under-prediction (college 12), but that is the case where the predicted and the actual GPAs are most nearly identical. A half of a standard deviation may impress some as a substantial overprediction.

Table 2 also supports and helps explain why investigators of the question of test bias have been willing to advocate the continued use of the SAT and other admissions tests with minority[3] students.

---

[3]Most investigations have dealt solely with black students as the minority and then the generalization has been extrapolated to other "minorities" (i.e., Mexican-Americans, the disadvantaged, low income females, etc.). One recent investigation, however (Cherdack, 1970), found that "the use of the L&S (Letters and Science) regression equation did not bias the predicted performance of the average student in each subgroup" (p. 141). The minorities he studied at two West Coast institutions were Educational Opportunity Program low-incomes, males, females, Negroes, Mexican-Americans, and Special Action students. This provides some evidence that the extrapolations may be justified by future studies.

Table 2

Predicted vs. Actual GPAs for Group 1 (Blacks)

School	(A) Predicted GPA	(B) Actual GPA	(A)-(B)	GPA SD	(A-B) /SD
1	2.77	2.58	.19	.48	.3958
2	2.74	2.30	.44	.58	.7586
3	2.53	2.26	.27	.58	.4655
4	2.35	2.14	.21	.57	.3684
5	2.43	2.06	.37	.66	.5606
6	2.33	2.17	.16	.65	.2462
7	2.60	2.15	.45	.66	.6818
8	2.47	2.45	.02	.73	.0274
9	2.31	2.11	.20	.64	.3125
10	2.43	1.96	.47	.65	.7231
11	2.35	2.30	.05	.55	.0909
12	2.10	2.11	-.01	.51	-.0196
13[a]	4.29	4.05	.24	1.30	.1846

[a]School 13 used seven levels of academic accomplishment rather than GPA. Seven was highest in our calculation.

A Validity Case Study Approach

Given that an admissions officer is impressed with the outcomes of the Gulliksen-Wilks tests (Table 1) and decides to make predictions using separate equations for blacks and whites, he could conduct a validity study that would provide the necessary information. Table 3 presents, for the institutions in this investigation, the information commonly obtained during such studies.

Table 3 may be examined in detail to see what generalizations from 13 case studies are supported.

There are a complex of possible generalizations from Table 3. If each institution is taken to represent a possible group of existing institutions, then the findings may support, among others, the following generalizations. For instance:

(1) Predictive validity is essentially the same for blacks and whites with

(a) moderate validity, or

(e.g., Colleges 1 and 3),

(b) high validity, or

(e.g., Colleges 5 and 12),

(c) special case (the test has no predictive validity for whites or blacks, e.g., College 7).

(2) The test has nonsignificant predictive validity for blacks, but does have moderate or high validity for whites (e.g., Colleges 2, 4, 9, 11, 13).

(3) Predictive validity exists for both groups, but is better for one than the other. For instance,

(a) better for blacks (e.g., College 10),

(b) better for whites (e.g., Colleges 8 and 6).

Table 3

Information on Group 1 (Blacks) and Group 2 (Whites) at

Thirteen Integrated Institutions

College	Group	Multiple Regression (V,M)	SAT-V M	SAT-V SD	SAT-M M	SAT-M SD	Multiple Regression Equations
1	1	.26	558	92	554	92	.0014V-.0002M+1.9154
	2	.27	648	73	660	69	.0013V+.0008M+1.5623
2	1	.15ns	517	92	510	101	.0000V+.0009M+1.8449
	2	.23	609	73	608	86	.0018V-.0003M+1.9635
3	1	.30	547	83	554	87	.0015V+.0009M+ .9547
	2	.38	692	59	696	65	.0018V+.0015M+ .7169
4	1	.18ns	565	75	538	73	.0014V-.0001M+1.3956
	2	.33	670	63	677	78	.0020V+.0010M+ .6830
5	1	.51	408	104	414	88	.0013V+.0027M+ .4225
	2	.55	545	99	562	98	.0026V+.0005M+1.1592
6	1	.25	528	69	516	78	.0018V+.0009M+ .7827
	2	.39	626	66	616	86	.0019V+.0022M+ .1882
7	1	.07ns	446	91	463	94	.0001V+.0004M+1.8913
	2	.15ns	604	87	623	94	.0004V+.0009M+2.0056
8	1	.27	409	68	426	65	.0027V+.0006M+1.0901
	2	.51	511	90	525	81	.0045V+.0012M+ .1137
9	1	.06ns	452	94	479	82	-.0002V+.0005M+1.9593
	2	.45	589	84	605	92	.0019V+.0014M+ .7867
10	1	.39	455	93	464	90	.0030V-.0007M+ .8926
	2	.25	600	75	617	60	.0014V+.0007M+1.4692
11	1	.08ns	473	75	489	94	.0006V+.0000M+2.0027
	2	.43	544	88	571	79	.0025V+.0007M+ .8199
12	1	.41	380	48	440	73	.0017V+.0024M+ .4106
	2	.49	446	80	498	73	.0035V+.0009M+ .3771
13	1	.15ns	624	91	590	96	.0014V+.0010M+2.5529
	2	.46	684	89	698	83	.0056V+.0007M+ .3847

Note: ns = nonsignificant. N equaled approximately 100 in each group except for schools 5 (140), 11 (69), and 12 (39 blacks, 100 whites).

Other investigators could perhaps support or propose additional generalizations using this same approach to the information in Table 3. This is left to their ingenuity. This study has demonstrated the urgency that institutions using the SAT (and other admissions tests and predictors) conduct validity analyses at their own institution over the short and the long haul. Assumptions about the differential validity or lack of it for blacks and whites should be routinely studied.

Summary

This study of test bias--differential prediction of freshman grade point average for blacks and whites--was essentially replicative, although some new considerations were presented. Thirteen integrated institutions, with increasing black student enrollment and SAT test scores as part of the admissions requirements, were studied by the method of comparisons of regression planes and predicted versus actual performance of black students.

The data tend to support the following generalizations:

(1) A single regression plane cannot be used to predict GPA for both blacks and whites in many of the institutions studied.

(2) If predictions of GPA from SAT scores are based upon regression equations suitable for majority students, then minority black students, as a group, are predicted to do about as well as (or better than) they actually do.

Examination of the validity study data from the 13 institutions demonstrated that a general conclusion applicable to all institutions is not justified. Admissions officers are urged to consider and do institutional self-studies routinely on the question of differential predictive validity.

References

Cherdack, A. N. The predictive validity of the Scholastic Aptitude Test for disadvantaged college students enrolled in a special education program. Unpublished doctoral dissertation, University of California at Los Angeles, 1970.

Chronicle staff. Federal survey of Negro undergraduate enrollments. Chronicle of Higher Education, April 21, 1969.

Cleary, T. A. Test bias: Prediction of grades of Negro and white students in integrated colleges. Journal of Educational Measurement, 1968, 5, 115-124.

Gulliksen, H., & Wilks, S. S. Regression tests for several samples. Psychometrika, 1950, 15, 91-114.

Kendrick, S. A., & Thomas, C. L. Transition from school to college. Review of Educational Research, 1970, 40(1), 151-179.

Linn, R. L., & Werts, C. E. Considerations for studies of test bias. Journal of Educational Measurement, 1971, 8(1), in press.

Potthoff, R. F. Statistical aspects of the problem of biases in psychological tests. Chapel Hill, N.C.: University of North Carolina at Chapel Hill, Department of Statistics, 1966 (Mimeo Series No. 479).

Rock, D. A. Motivation, moderators, and test bias. Toledo Law Review, 1970, in press.

Stanley, J. C., & Porter, A. C. Correlation of Scholastic Aptitude Test scores with college grades for Negroes versus whites. Journal of Educational Measurement, 1967, 4, 199-218.

APPENDIX

In some institutions, because there seems to be no significant relationship between black student performance and test scores, an admissions officer might explore whether to predict all future black applicants with the mean of the past year's black student group. This would be based upon the correct assumption that the mean is the best prediction in the absence of information. (Of course, this prediction does not provide any differential information for purposes of selection or guidance of the black applicants.) If the admissions officer does explore this question, he may well ask, eventually: What is the lowest composite SAT score that I could have entered into our majority regression equation that would have yielded the same mean predicted score? He asks this because he has become concerned that in the past he has been using the majority regression equation without knowledge of the lack of relationship. If he knows the lowest composite score that would have yielded a prediction of the black mean or better, he has a rather good index to either the benefit or injustice of his prediction system for black applicants as a group.

In Table A a calculation of this lower limit value for the case of the SAT V+M composite is presented for the five institutions where the validity for Group 1 (blacks) did not reach significance.

It is apparent from Table A that in institutions like College 2, there is no real danger of discriminating against blacks with test scores entered in a majority regression equation. The equation predicts higher performance for any value of V+M than absence of the information would allow one to predict.

Table A

College	Using majority regression equation	Lowest score[a] (V+M) yielding prediction equal to black mean GPA	Percent blacks at or above lowest limit	Mean black composite score
2	Y= .0006(V+M)+ 2.1212	(298)= below possible score (V+M)	100	1027
4	Y= .0014(V+M)+ .7832	1055	63 1/2	1102
9	Y= .0017(V+M)+ .7992	771	87	931
11	Y= .0017(V+M)+ .7150	932	63 1/4	962
13	Y= .0033(V+M)+ .2182	1161	62 1/2	1214

[a]This value is calculable by solving the zero-order regression equation for the predictor value rather than the GPA. The restated equation used was:

$$(V+M) = \frac{\text{black actual GPA - white intercept}}{\text{white beta weight}}$$

In the four other institutions and those like them, to the extent that more than 50 percent of the black group are at or above the lower limit calculated, there has been a tendency toward overprediction of their performance. This tendency was noted in the body of this report. It is also possible to examine in detail black students with below the lower limit composite scores reported in Table A. These students would be predicted to perform lower than the mean of their group if the majority regression equation was used. Therefore, to the extent that more than 50 percent of them exceed the mean of their respective black group it would indicate a tendency to underpredict this subgroup of the black minority group. Table B reports data on this subgroup at the same five institutions.

Table B indicates a tendency toward underprediction at only one institution, College 11. The other four institutions show no such tendency, and, in fact, College 4 varies as much toward overprediction as College 11 does toward underprediction; in both cases the tendency is slight compared to the percentage difference above 50 percent in Table A for those students above the lower limit composite score.

Discussion

This appendix has reported validity study information addressed to only one concern of admissions officers using test data in prediction of applicant performance. It did not address itself to those more numerous admission situations where there are differential but significant validities for black and white students or the problems of considering other information along with test information in decisions for admission. However, the data reported may illustrate the complexity and necessity for further careful work in this area without either rejecting or continuing test usage uncritically.

Table B

College	(1) N of black group	(2) N of black subgroup below composite score in Table A	Percent of (2) above black mean GPA
2	98	0	NA
4	92	34	44
9	100	13	46
11	68	25	56
13	104	39	50[a]

[a]School 13 was adjusted for use of levels of performance rather than GPA.